A Document as a Small World

Yutaka Matsuo† Yukio Ohsawa‡ Mitsuru Ishizuka†

†Graduate School of Engineering
University of Tokyo
Tokyo 113-8656, Japan

‡Graduate School of Systems Management
University of Tsukuba
Tokyo 112-0012, Japan

Abstract

The small world topology is known widespread in biological, social and man-made systems. This paper shows that the small world structure also exists in documents, such as papers. A document is represented by a network; the nodes represent terms, and the edges represent the co-occurrence of terms. This network is shown to have the characteristics of being small world, i.e., highly clustered and short path length. Based on the topology, we can extract important terms, even if they are rare, by measuring their contribution to the graph being small world.

1 Introduction

Graphs that occur in many biological, social and man-made systems are often neither completely regular nor completely random, but have instead a "small world" topology in which nodes are highly clustered yet the path length between them is small [10, 8]. For instance, if you are introduced to someone at a party in a small world, you can usually find a short chain of mutual acquaintances that connects you together. In the 1960s, Stanley Milgram's pioneering work on the small world problem showed that any two randomly chosen individuals in the United States are linked by a chain of six or fewer first-name acquaintances, known as "six degrees of separation" [4]. Watts and Strogatz have shown that a social graph (the collaboration graph of actors in feature films), a biological graph (the neural network of the nematode worm *C. elegans*), and a man-made graph (the electrical power grid of the western United States) all have a small world topology [10, 9]. World Wide Web also forms a small world network [1].

In the context of document indexing, an innovative algorithm called *KeyGraph* [5] is developed, which utilizes the structure of the document. A document is represented as a graph, each node corresponds to a term[1] , and each edge corresponds to the co-occurrence of terms. Based on the segmentation of this graph into clusters, *KeyGraph* finds keywords by selecting the term which co-occurs in multiple clusters. Recently, *KeyGraph* has been applied to several domains, from earthquake sequences [6] to register transaction data of retail stores, and showed remarkable versatility.

[1] A term is a word or a word sequence.

In this paper, inspired by both small world phenomenon and *KeyGraph*, we develop a new algorithm to find important terms. We show at first the graph derived from a document has the small world characteristics. To extract important terms, we find those terms which contribute to the world being small. The contribution is quantitatively measured by the difference of "small-worldliness" with and without the term.

The rest of the paper is organized as follows. In the following section, we first detail the small world topology, and show that some documents actually have small world characteristics. Then we explain how to extract the important terms in Section 3. Finally, we discuss future works and conclude this paper.

2 Term Co-occurrence Graph and Small World

2.1 Small-worldliness

We treat an *undirected*, *unweighted*, *simple*, *sparse* and *connected* graph. (We expand to an *unconnected* graph in Section 3.) To formalize the notion of a small world, Watts and Strogatz define the clustering coefficient and the characteristic path length [10, 9]:

- The *characteristic path length*, L, is the path length averaged over all pairs of nodes. The path length $d(i, j)$ is the number of edges in the shortest path between nodes i and j.

- The *clustering coefficient* is a measure of the cliqueness of the local neighbourhoods. For a node with k neighbours, then at most ${}_kC_2 = k(k-1)/2$ edges can exist between them. The

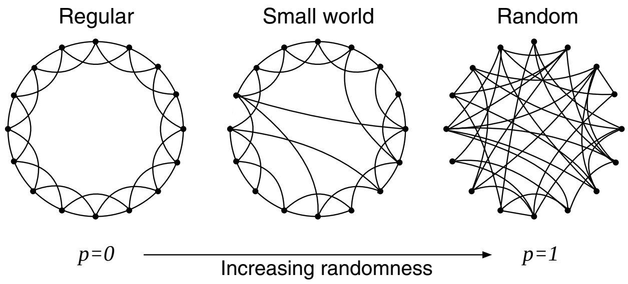

Figure 1: Random rewiring of a regular ring lattice.

clustering of a node is the fraction of these allowable edges that occur. The clustering coefficient, C is the average clustering over all the nodes in the graph.

Watts and Strogatz define a small world graph as one in which $L \geq L_{rand}$ (or $L \approx L_{rand}$) and $C \gg C_{rand}$ where L_{rand} and C_{rand} are the characteristic path length and clustering coefficient of a random graph with the same number of nodes and edges. They propose several models of graphs, one of which is called β-Graphs. Starting from a regular graph, they introduce disorder into the graph by randomly rewiring each edge with probability p as shown in Fig.1. If $p = 0$ then the graph is completely regular and ordered. If $p = 1$ then the graph is completely random and disordered. Intermediate values of p give graphs that are neither completely regular nor completely disordered. They are small worlds.

Walsh defines the proximity ratio

$$\mu = (C/L) \,/\, (C_{rand}/L_{rand}) \tag{1}$$

as the small-worldliness of the graph [8]. As p increases from 0, L drops sharply since a few long-range edges introduce short cuts into the graph. These short cuts have little effect on C. As a consequence the proximity ratio μ rises rapidly and the graph develops a small world topology. As p approaches 1, the neighbourhood clustering start to break down, and the short cuts no longer have a dramatic effect at linking up nodes. C and μ therefore drop, and the graph loses its small world topology. In Table 1, we can see μ is large in the graphs with a small world topology.

In short, small world networks are characterized by the distinctive combination of high clustering with short characteristic path length.

2.2 Term Co-occurrence Graph

A graph is constructed from a document as follows. We first preprocess the document by stemming and removing *stop words*, as in [7]. We apply n-gram to count phrase frequency. Then we regard the title of the document, each section title and each caption of figures and tables as a sentence, and exclude all the figures, tables, and references. We get a list of sentences, each of which consists of words (or phrases). In other words, we get a basket data where each item is a term, discarding the information of term orderings and document structures.

Then we pick up *frequent terms* which appear over a user-given threshold, f_{thre} times, and fix them as nodes. For every pair of terms, we count the *co-occurrence* for every sentences, and add an edge if the Jaccard coefficient exceeds a threshold, J_{thre}[2]. The Jaccard coefficient is simply the number of sentences that contain both terms divided by the number of sentences that contain either terms. This idea is also used in constructing a referral network from WWW pages [3]. We assume the length of each edge is 1.

Table 2 is statistics of the small-worldliness of 57

Table 1: Characteristic path lengths L, clustering coefficients C and proximity ratios μ for graphs with a small world topology [8] (studied in [10])).

	L	L_{rand}	C	C_{rand}	μ
Film actor	3.65	2.99	0.79	0.00027	2396
Power grid	18.7	12.4	0.080	0.005	10.61
C. elegans	2.65	2.55	0.28	0.05	4.755

The graphs are defined as follows. For the film actors, two actors are joined by an edge if they have acted in a film together. For the power grid, nodes represent generators, transformers and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction.

Table 2: Statistical data on proximity ratios μ for 57 graphs of papers in WWW9.

	L	L_{rand}	C	C_{rand}	μ
Max.	4.99	3.58	0.38	0.012	22.81
Ave.	5.36	—	0.33	—	15.31
Min.	8.13	2.94	0.31	0.027	4.20

We set $f_{thre} = 3$. We restrict attention to the giant connected component of the graph, which include 89% of the nodes on average. We exclude three papers, where the giant connected component covers less than 50% of the nodes. We don't show the L_{rand} and C_{rand} for the average case, because n and k differs dependent on the target paper. On average, $n = 275$ and $k = 5.04$.

[2] In this paper, we set J_{thre} so that the number of neighbors, k, is around 4.5 on average.

graphs, each constructed from a technical paper that appeared at the 9th international World Wide Web conference (WWW9) 2000 [11]. From this result, we can conjecture these papers certainly have small world structures. However, depending on the paper, the small-worldliness varies.

One reason why the paper has a small world structure can be considered that the author may mention some concepts step by step (making the clustering of related terms), and then try to merge the concepts and build up new ideas (making a 'shortcut' of clusters). The author will keep in mind that the new idea is steadily connected to the fundamental concepts, but not redundantly. However, as we have seen, the small-worldliness varies from paper to paper. Certainly it depends on the subject, the aim, and the author's writing style of the paper.

3 Finding Important Terms

3.1 Shortcut and Contractor

Admitting that a document is a small world, how does it benefit us? We try here to estimate the importance of a term, and pick up important terms, though they are rare in the document, based on the small world structure. We consider 'important terms' as the terms which reflect the main topic, the author's idea, and the fundamental concepts of the document.

First we introduce the notion of a *shortcut* and a *contractor*, following the definition in [9].

Definition 3.1
The range $R(i,j)$ is the length of the shortest path between i and j in the *absence* of that edge. If $R(i,j) > 2$, then the edge (i,j) is called a *shortcut*.

Applying the notion of "shortcuts" in terms of nodes, we can get the definition of "contractor."

Definition 3.2
If two nodes u and w are both elements of the same neighbourhood $\Gamma(v)$, and the shortest path length between them that does not involve any edges adjacent with v is denoted $d_v(u,w) > 2$, then v is said to *contract* u and w, and v is called a *contractor*.

In our first thought, if $d_v(u,w)$ is large, the corresponding term of contractor v might be interesting, because they bridge the distant notions which rarely appear together. However, such a node sometimes connects the nodes far from the center of the graph, i.e. the main topic of the document. Below we take into account the whole structure of the graph, calculating the contribution of a node to make the world small.

To treat the disconnected graph, we expand the definition of path length (though Watts restricts attention to the giant connected component of the graph).

Definition 3.3
An *extended* path length $d'(i,j)$ of node i and j is defined as follows.

$$d'(i,j) = \begin{cases} d(i,j), & \text{if } (i,j) \text{ are connected,} \\ w_{sum}, & \text{otherwise.} \end{cases} \tag{2}$$

where w_{sum} is a constant, the sum of the widths of all the disconnected subgraphs. $d(i,j)$ is a path length of the shortest path between i and j in a connected graph.

If some edges are added to the graph and some parts of the graph gets connected, $d'(i,j)$ will not increase, unless the length of an edge is negative. Thus $d'(i,j)$ is one of the upper bounds of the path length considering the edges will be added.

Definition 3.4
Extended characteristic path length L' is an extended path length averaged over all pairs of nodes.

Definition 3.5
L'_v is an extended path length averaged over all pairs of nodes except node v. L'_{G_v} is the extended characteristic path length of the graph without node v.

In other words, L'_v is the characteristic path length regarding the node v as a corridor (i.e., a set of edges). For example, if v is neighboring u, w, and z, then $(u,w), (u,z), and (w,z)$ are considered to be linked. And L'_{G_v} is the extended characteristic path length assuming the corridor doesn't exist.

Definition 3.6
The *contribution*, CB_v, of the node v to make the world small is defined as follows.

$$CB_v = L'_{G_v} - L'_v \tag{3}$$

We don't pay attention to the clustering coefficient, because adding or eliminating one node affects the clustering coefficient little.

If node v with large CB_v is absent in the graph, the graph gets very large. In the context of documents, the topics are divided. We assume such a term help merge the structure of the document, thus important.

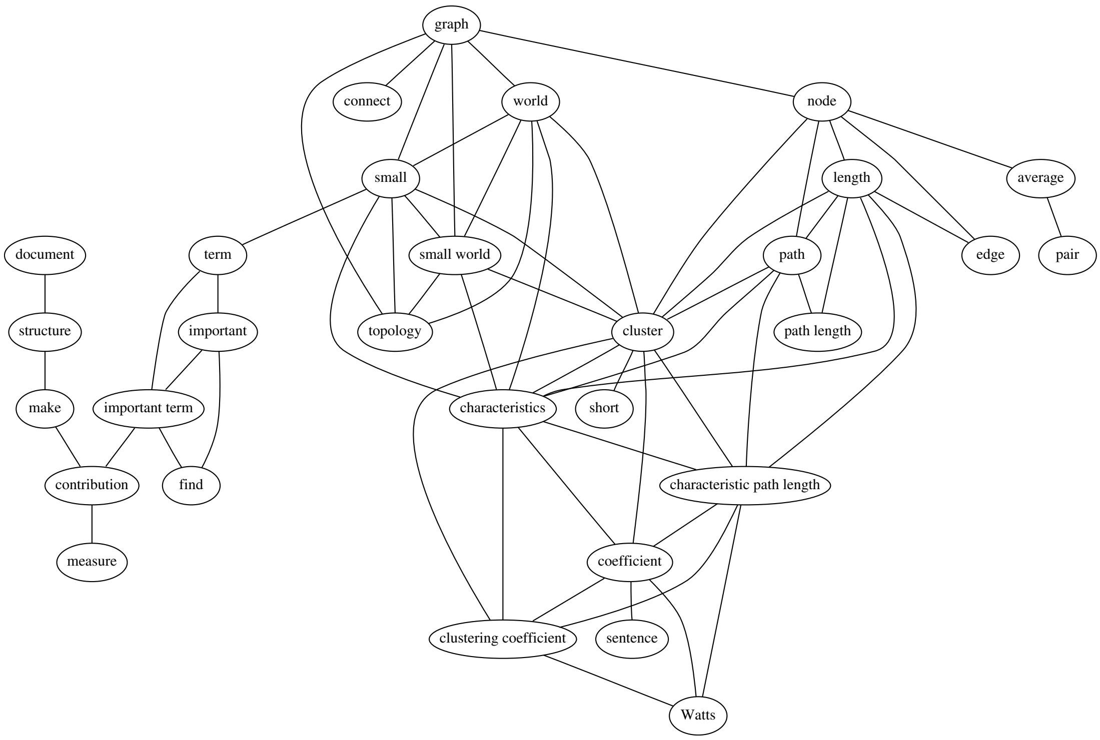

Figure 2: Small World of This Paper.

Table 3: Frequent terms in this paper.

Term	Frequency
graph	39
small	37
world	37
term	34
small world	30
node	29
paper	21
length	21
document	19
edge	19

Table 4: Terms with 10 largest CB_v in this paper.

Term	CB_v	Frequency
small	3.05	37
term	2.80	34
important term	1.93	7
contribution	1.64	6
node	1.00	29
make	0.82	6
cluster	0.57	15
graph	0.54	39
coefficient	0.52	8
average	0.50	8

3.2 Example

We show the example experimented on this paper, i.e., the one you are reading now[3] . Table 3 shows the frequent terms and Table 4 shows the important terms measured by CB_v. Comparing two tables, the list of important terms includes the author's idea, e.g., *important term* and *contribution*, as well as the important basic concept, e.g., *cluster* and *coefficient*, although they are rare terms. However the list of frequent terms simply show the components of the papers, and are not of interest.

We can also measure the contribution of an edge, CB_e, to make the world small, defined similary as CB_v. However, if we look at the pairs of terms in

[3] We ignore the effect of *self-reference*; it's sufficiently small.

Table 5: Pairs of Terms with 10 Largest CB_e.

Pair	CB_e
small – term	3.07
important term – contribution	1.93
make – contribution	1.22
node – average	0.98
structure – make	0.82
cluster – short	0.55
graph – connect	0.53
coefficient – sentence	0.52
average – pair	0.49
contribution – measure	0.48

Table 5, it is hard to understand what they suggest. There are numbers of relations between two terms, so we cannot imagine the relation of the pairs right away.

Lastly, Fig. 2 shows the graphical visualization of the world of this paper. (Only the giant connected component of the graph is shown, though other parts of the graph is also used for calculation.) We can easily point out the terms without which the world will be separated, say *small* and *important term*.

4 Conclusion

Watts mentions in [9] the possible applications of small world research, including "the train of thought followed in a conversation or succession of ideas leading to a scientific breakthrough." In this paper, we have focused on the papers rather than conversation or succession of ideas. The future direction of our research is to treat *directed* or *weighted* graph for finer analyses of the document.

We expect our approach is effective not only to document indexing, but also to other graphical representations. To find out structurally important parts may bring us deeper understandings of the graph, new perspectives, and chances to utilize it. We are interested in a big structural change caused by a small change of the graph. The importance of weak ties, which is a short cut between clusters of people, was mentioned 30 years ago [2]. A change, which makes the world very small, may sometimes be very important.

References

[1] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the World Wide Web. *Nature*, 401, 1999.

[2] M. Granovetter. Strength of weak ties. *American Journal of Sociology*, (78):1360–1380, 1973.

[3] H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magagine*, 18(2), 1997.

[4] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Technical Report TR 99-1776, Cornell University, 1999.

[5] Y. Ohsawa, N. E. Benson, and M. Yachida. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proc. Advanced Digital Library Conference (IEEE ADL'98)*, 1998.

[6] Y. Ohsawa and M. Yachida. Discover risky active faults by indexing an earthquake sequence. In *Proc. Discovery Science*, pages 208–219, 1999.

[7] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1988.

[8] T. Walsh. Search in a small world. In *Proc. IJCAI-99*, pages 1172–1177, 1999.

[9] D. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton, 1999.

[10] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 1998.

[11] http://www9.org/.